REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	~~April 10, 1998~~	Final Report (4/1/94-3/31/98)

4. TITLE AND SUBTITLE
Bootstrap Calibration, Model Selection and Tree-Structured Methods

5. FUNDING NUMBERS
DAAH04-94-G-0042

6. AUTHOR(S)
Wei-Yin Loh

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
University of Wisconsin-Madison
750 University Avenue
Madison, WI 53706

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARO 32330.13-MA

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Several problems in variable selection and decision trees were solved. In the case of linear regression models with increasing number of covariates, a method based on ordering the covariates in terms of their t-statistics is shown to be asymptotically consistent as the sample size increases. This result holds for the fixed design situation as well as that of random covariates.

A new unbiased method of split selection for classification trees was developed and implemented into computer software. The method is unbiased in the sense that when all the covariates are unrelated to the response variable, each covariate has an equal chance of being selected to split a node. No previous algorithm has this property. Bootstrap calibration plays a critical role in the algorithm. Empirical evaluations of the algorithm show that it is as accurate as the best classifiers from the statistical and computer science literature. It has the additional benefit of being one of the fastest algorithms.

14. SUBJECT TERMS
Bootstrap, Recursive Partitioning, Decision Trees

15. NUMBER IF PAGES
6

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

BOOTSTRAP CALIBRATION, MODEL SELECTION AND TREE-STRUCTURED METHODS

FINAL REPORT

WEI-YIN LOH

APRIL 6, 1998

U. S. ARMY RESEARCH OFFICE
GRANT NUMBER DAAH04-94-G-0042
UNIVERSITY OF WISCONSIN, MADISON

1 Problems studied

1. Tests of equality of variances.
2. Variable selection for linear models with high-dimensional covariates.
3. Split selection methods for classification trees.
4. Comparison of decision trees and other classification methods.
5. Unbiased piecewise-linear regression trees.

2 Summary of important results

The following results were obtained for each of the problems listed above. References refer to the list of publications in Section 3.

1. Seven tests of equality of variances were compared in terms of robustness and power in a simulation experiment with small to moderate sample sizes. The data were assumed to come from a location-scale family with unknown means, variances, and density functions. The tests considered were the Levene test, the Bartlett test with and without kurtosis adjustment, the Box-Andersen test, and three jackknife tests. The bootstrap versions of these tests were also compared. It is found that the Levene test and one jackknife test, as well as the bootstrap versions of the Levene test, the Bartlett test with kurtosis adjustment, and two jackknife tests, are robust. Among these, the bootstrap version of the Levene test tends to have the highest power. The results are published in [7].

2. The problem is that of variable selection in linear regression models when the number of covariates is allowed to increase with the sample size. The approach in [5] for the fixed design situation is extended to the case of random covariates. This yields a unified consistent selection criterion for both random and fixed covariates. By using t-statistics to order the covariates, the method requires much less computation than an all-subsets search. The method can be applied to autoregressive model selection with increasing order. Simulation experiments were carried out to validate the theory. The results are published in [8].

3. Classification trees based on exhaustive search algorithms (such as AID and CART) tend to be biased towards selecting variables that allow more splits. As a result, such trees need to be interpreted with caution. An algorithm called QUEST that has negligible selection bias was developed. Its split selection strategy shares similarities with the FACT method, but it yields binary splits and the final tree can be selected by a direct stopping rule or by pruning. Real and simulated data were used to compare QUEST with the exhaustive search approach. QUEST is shown to be substantially faster and the size and classification accuracy of its trees are typically comparable to those of exhaustive search. The results are reported in [9]. Compiled executable versions of the computer program are available for downloading from the PI's homepage (`http://www.stat.wisc.edu/~loh/`). The QUEST algorithm has been adopted by the commercial software publishers of SPSS and STATISTICA for inclusion in their packages.

4. Twenty two decision tree, nine statistical, and two neural network classifiers were compared on thirty-two datasets in terms of classification error rate, computational time, and (in the case of trees) number of terminal nodes. It is found that the average error rates for a majority of the classifiers are not statistically significant but the computational times of the classifiers differ over a wide range. The statistical classifier POLYCLASS based on a logistic regression spline algorithm has the lowest average error rate. However, it is also one of the most computationally intensive. The classifier based on standard polytomous logistic regression and the QUEST classification tree with linear splits have the second lowest average error rates but are about 50 times faster than POLYCLASS. Among decision tree classifiers with univariate splits, the classifiers based on the C4.5, IND-CART, and QUEST algorithms have the best combination of error rate and speed, although the C4.5 trees tend to have about twice as many nodes as those from the other two algorithms. The C4.5 classifier based on rules also has good accuracy, but it does not scale as well as the other methods. These results are reported in [11].

5. A piecewise-constant regression tree model can be valuable for the insights that its tree structure provides. However, the standard exhaustive search approach to tree construction has three weaknesses that

limits its usefulness. First, it possesses a variable selection bias that can lead to erroneous conclusions. Second, the piecewise-constant trees tend to have many levels of splits, which hinder interpretation. Third, its split selection criterion focuses only on one predictor variable at a time. As a result, it may fail to detect interactions between two predictors, or require more than one split to uncover them.

An alternative approach, called GUIDE, to tree construction is developed that (1) employs significance tests and the bootstrap to correct for biases in variable selection, (2) permits the fitting of piecewise-linear models to reduce tree complexity, and (3) chooses splits according to measures of curvature within individual predictors as well as interactions between pairs of predictors. The method accepts ordered and unordered predictor variables, with unordered variables being allowed to split the nodes but not participate in the linear model equations. Simulation experiments show that the selection bias of the exhaustive search approach can be quite severe. They also show that GUIDE is effective in correcting the bias. The algorithm and results are reported in [12].

3 Publications and manuscripts submitted for publication

1. Tree-structured proportional hazards regression modeling (with H. Ahn). *Biometrics*, **50**, 471–485, 1994.

2. Piecewise-polynomial regression trees (with P. Chaudhuri, M.-C. Huang and R. Yao). *Statistica Sinica*, **4**, 143–167, 1994.

3. Bias and variance reduction in estimation of model dimension (with X. Zheng). *Proceedings of the American Mathematical Society.* **122**, 1263–1272, 1994.

4. Generalized regression trees (with P. Chaudhuri, W.-D. Lo and C.-C. Yang). *Statistica Sinica*, **5**, 641–666, 1995.

5. Consistent variable selection in linear models (with X. Zheng). *Journal of the American Statistical Association*, **90**, 151–156, 1995.

6. Bootstrapping binomial confidence intervals (with X. Zheng). *Journal of Statistical Planning and Inference*, **43**, 355–380, 1995.

7. A comparison of tests of equality of variances (with T.-S. Lim). *Computational Statistics and Data Analysis*, **22**, 287–301, 1996.

8. A consistent variable selection criterion for linear models with high-dimensional covariates (with X. Zheng). *Statistica Sinica*, 7, 311–325, 1997.

9. Split selection methods for classification trees (with Y.-S. Shih). *Statistica Sinica*, 7, 815–840, 1997.

10. Asymptotic theory for Box-Cox transformations in linear models (with K. Cho, R. A. Johnson and I. Yeo). Submitted to *Annals of Statistics.*

11. An empirical comparison of decision trees and other classification methods (with T.-S. Lim and Y.-S. Shih). Submitted to *Machine Learning.*

12. Unbiased regression trees. Submitted to *Journal of the American Statistical Association.*

4 List of participating scientific personnel

The grant provided research assistantship support to the following four PhD students.

1. Yunfei Chen (Research assistant, PhD expected summer 1999)

2. Kwanho Cho (Research assistant, PhD expected summer 1998)

3. Hyunjoong Kim (Research assistant, PhD expected summer 1998)

4. Tjen-Sien Lim (Research assistant, PhD expected summer 1998)

The following PhD students received their degrees under the PI's supervision during the grant period. The thesis titles are given in italics.

1. Peng Qu, 1994. *Application of Box-Cox transformations to discrimination for the two-class problem.*

2. Ruji Yao, 1994. *Regression trees.*

3. Xujie Yu, 1994. *Analysis of contingency tables.*

4. Xiaodong Zheng, 1994. *Bootstrap theory and methods.*

5. Chongqing Yan, 1995. *Regression trees and nonlinear time series modeling.*